Estimation Of Air Quality Index In Delhi By Merging Neural Networks And Multiple Regression Techniques with Principal Components Analysis

Shriniketan Kulkarni
Electrical and Electronics Engineering Department
Manipal Institute of Technology, MAHE
Udupi, India
Kulkarnishriniketan27@gmail.com

Harneet Singh Bali
Electrical and Electronics Engineering Department
Manipal Institute of Technology, MAHE
Udaipur, India
harneet2512singh@gmail.com

Rajashree Krishna
Computer Science Engineering Department
Manipal Institute of Technology, MAHE
Udupi, India
raji.krish@manipal.edu

***Abstract*—A major focus of air quality research in recent years has been the AQI measurement as a way to gauge the harm pollution does to people's health and well-being in cities. Air Quality Index (AQI) accuracy is the primary goal of this research, which uses PCA and Artificial Neural Network (ANN) approaches. Our investigators are looking into the city of Delhi. To forecast the air quality index, The main components score (PCS) of 11 historical air quality and meteorological data is used in an ANN model (AQI). Delhi is the subject of this investigation. In order to make accurate forecasts of the air quality index, ANN models make use of the main components score (PCS) of 11 meteorological and historical air quality indicators (AQI). A comparison is made between ANN and MLR models, which are commonly used to estimate the AQI. Other than PCA, you may also reduce the eleven parameters to just eight PCs. PC-ANN (PC-ANN) models use the eight PCs as input data. The R2, RMSE, MAPE, and MAE values were used to make comparisons between the various models and hypotheses. The PC-ANN model outperforms all others when considering the complexities of air pollution. As a result, the PC-ANN approach may be utilized to make better decisions and address atmospheric management challenges.**

***Index Terms*—Air Quality Index, Neural Networks, Multiple Regression, Principal Component Analysis (PCA)**

I. Introduction

India's air pollution situation is growing more acute, especially in the capital city of Delhi. Power plants, industry, domestic heating, and fuel-burning autos, as well as natural calamities, are all common sources of global warming. A major effect of air pollution on people's health, especially in cities, is a respiratory disease and other chronic health conditions. Long-term climate consequences of global warming and the greenhouse effect can be seen. This indicates that perhaps the air we breathe is not pure, but rather polluted, as it contains several hazardous substances and particles that are harmful to human health. For many years, research has shown that poor air quality has negative consequences on human health, including cardiovascular disease[1], genetic childhood asthma, and more recently, neurological problems[2]. The National Ambient Air Monitoring Network collects data on the concentrations of various contaminants in the air, however, this data is difficult to comprehend for the general public. The Central Pollution Control Board (CPCB) creates an AQI for Indian cities on a national level. The AQI (Air Quality Index) is a measure of air pollution in a specific location. As a result, AQI measures the real quality of the air we breathe, and this, in turn, has been shown to have a wide range of implications for our health. The concentrations of distinct air pollutants in various types of buildings, including residential, commercial, and industrial, are used to calculate an air quality index (AQI). Monitoring data is compiled and standardized into a single index in a number of different ways. Indexing and descriptors for air pollution are therefore highly variable amongst different regions and countries. For the public, air quality indicator data provides a simple way to monitor air quality in their area, region, or country without having to know all of its complexities.

II. Literature Review

Anikender Kumar and PramilaGoyal [3] published research that predictions the daily AQI value for Delhi, India utilizing the prior record of AQI and climatic data using Principal Component Regression (PCR) and Multiple Linear Regression techniques. With data from 2000-2005 and several equations, they are able to predict the daily AQI for 2006 based on prior records. The Multiple Linear Regression Technique was used to compare this projected AQI to an observed AQI in 2006 during the monsoon, winter, summer, and post-monsoon seasons. With the PCA method, the correlations between the independent variables can be discovered. There were fewer

predictor variables and no collinearity between predictor variables in multiple linear regression when the Principal Component method was used. In the winter, the Principal Component Regression performs better than any other season for forecasting the AQI. When predicting the future AQI, this research relied only on meteorological indicators, excluding possible health-harming ambient air pollutants.

Researchers Eman Sarwat and Ghada I. El-Shanshoury [5] conducted a research in which they used PCA and ANN approaches to accurately estimate the Air Quality Index (AQI). For the year 2014, Ain Sokhna city will be the focus of this research. Computer-aided analysis and artificial neural networks are used to estimate the air quality index using 10 historical air quality and weather indicators. The feed-forward ANN model has a higher R2 and lower error rates than the MLR model when 10 original parameters are used as inputs (Method 1). PC-ANNs utilizing the Varimax technique have a superior R2 value and error rate than Method l (PC-ANNs without the Varimax method). The outcomes of Varimax PC-ANN prediction and actual values are also virtually comparable. Although the accuracy of Methods 3 and 4 (PC-ANNs applying Equamax and Quartimax methodologies) is lesser than that of Methods 1 and 2, these methods can nevertheless predict AQI values reliably. This is in comparison to MLR and PCR models, whose accuracy is higher. Varimax rotated PC scores and the ANN model were shown to be more accurate and efficient in predicting AQI outcomes. In addition, using ANN or PC-ANN models rather than MLR or PCR improves the accuracy of estimates. When making decisions and dealing with difficulties related to better atmospheric management in the local area, the PC-ANN and ANN models are particularly valuable tools.

III. Methodology

A. Air Pollutants Data

Average daily air quality statistics for NSIT Dwarka for the pollutants sulphur dioxide (SO_2), nitrogen dioxide (NO_2), Nitrogen Monoxide (NO), ozone (O_3), benzene, toluene and respirable particulate matter ($PM_{2.5}$) from 2013 to 2021 (One of the well known and prominent places in New Delhi). A variety of environmental parameters, like relative humidity (RH), wind speed (WS), wind direction (WD), and sun radiation (SR), were also taken into account for determining the levels of these pollutants in Delhi's central pollution control board (CPCB). PM_{10} is also one of the major air pollutants criteria which is not considered in this study as sufficient data of PM_{10} was not available for the period 2013 – 2021.

B. Meteorological Data

The daily average surface meteorological data includes the following variables: temperature (T), maximum temperature (TM), and minimum temperature (TM), atmospheric pressure at sea level (P), average relative humidity (H), total rainfall and snowmelt, visibility (VV), average wind speed, and maximum sustained wind speed. For the years 2013–2021, Delhi's Safdar-jung Airport Station, operated by the Indian Meteorological Department (IMD), has provided weather information. Safdarjung airport is about 20 kms away from NSIT Dwarka this is the closest station we could find

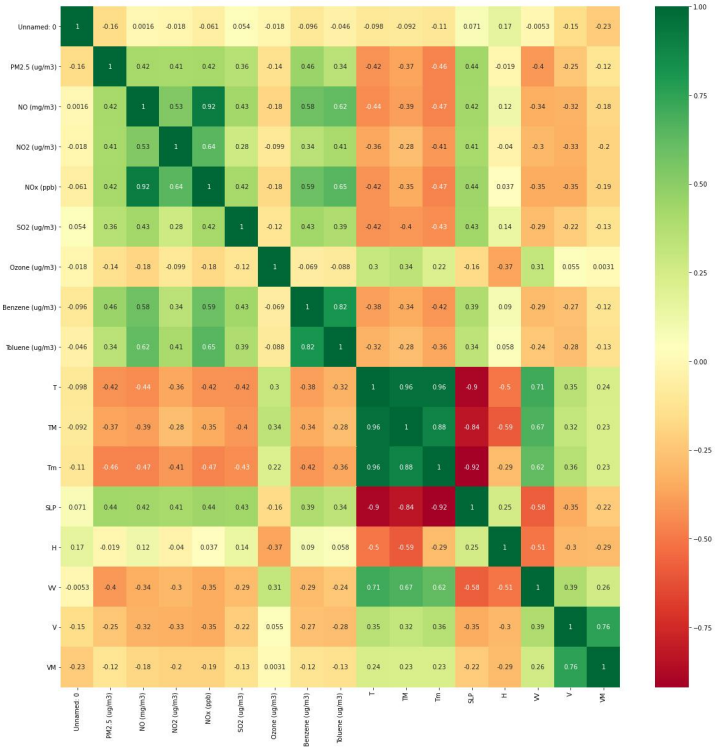

Fig. 1. Correlation Heat map

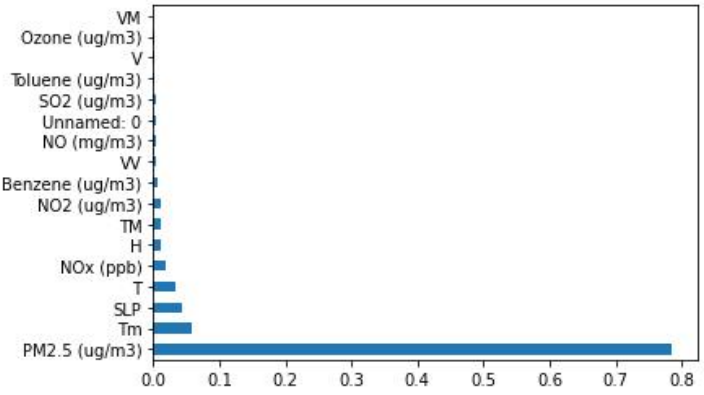

Fig. 2. Feature Importance based on correlation coefficients

C. AQI Calculation

The first two processes in computing the AQI are to create sub-indices for each pollutant and to aggregate the sub-indices. Epidemiological research and the Indian NAAQS, which both demonstrate that certain pollutants cause a health risk, are used to estimate the break-point concentration for each pollutant. A variety of literature-reported break-point concentrations and air quality criteria have been identified (EPA, 1999). The table

below illustrates a variety of index values that may be used in India to measure air quality and how it affects people's health.

TABLE I
AQI CATEGORY FOR OZONE AND PARTICLE POLLUTION

For this AQI.	use this descriptor.	and this color
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

All $PM_{2.5}$,NO_2, and NO values are expressed in (ug/m3). The air quality is good, although a small number of individuals may have a moderate health risk due to certain toxins. Moderate: Individuals who are members of vulnerable groups might suffer from negative health effects. Poor: People who belong to marginalized groups are more likely to become ill. Everyone is at risk for more serious health implications if the situation is considered to be very poor. Extremely serious: health warnings and emergency alerts are triggered. The formula of AQI is given:

$$I_{\mathrm{p}} = \left[\frac{(I_{\mathrm{Hi}} - I_{\mathrm{Lo}})}{(\mathrm{BP}_{\mathrm{Hi}} - \mathrm{BP}_{\mathrm{Lo}})}\right](C_{\mathrm{P}} - \mathrm{BP}_{\mathrm{Lo}}) + I_{\mathrm{Lo}} \qquad (1)$$

where I_p denotes the air quality index and C_p is actual ambient concentration of pollutant 'p.' For each of these breakpoints, a sub-index value is assigned: I_{Hi} for BP_{Hi}, and I_{Lo} for BP_{Lo}. BP_{Hi}'s sub index value is equal to or greater than C_p's sub index value. As part of the statistical models, each pollutant's AQI has been obtained separately, and that AQI of the day ((NO, NO_2, $PM_{2.5}$, and SO_2) is used as one of the input factors for each of the statistical models.

D. Principal Component Analysis (PCA)

Performing a factor analysis can be accomplished through Principal Component Analysis (PCA). Among the feed-forward networks investigated in this study was the Multilayer Perceptron (MLP), which can use backpropagation, conjugate gradient, and other techniques to improve its accuracy and speed. Linear combinations of the original measurements provide the basis of the new set of variables. As a result of the linear combinations, each composite variate accounts for less of the total variation than the one before it. Each component's variance will vary in accordance with its relative importance, starting with the most important one (the first component). Only the top few principal components (or, the eigenvector-eigenvalue pairs) should be accounted for more than 60% of the total variance when calculating the overall Air Quality Index (PCs). A small fraction of the total variance is explained by higher-order PCs, which therefore are regarded as noise. In our case to account for the maximum variation in the data we had to select the first 8 Principal Components (PCs).

E. Multiple Linear Regression

In atmospheric modeling, multiple linear regression (MLR) is frequently used. [7]. With the help of a linear equation and an estimation of the percentage contribution each parameter has made to air pollution, this method has been utilized for studying the link between numerous independent and dependent variables. Air quality index (AQI) data is used as a dependent variable in this research to support the relationship between AQI and meteorological factors (11 parameters in our example). The model is obtained using the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad (2)$$

where, y is the response variable, x_1, x_2, x_n are the explanatory variable,β_0, β_1, β_2, …, β_n are regression coefficients.

F. Principal Component Regression

PCA (Principal Component Analysis) is combined with OLS in Principal Component Regression (ordinary least square technique). As part of principle components regression, PCA is used to choose a subset of the orthogonal x-variable principal components (the principal components) as the variables to be utilized in forecasting y. Identifying the principal components of a linear regression model fitted using the classic least squares approach is the primary goal of doing principal component analysis (PCR).PCR can be used to compress a large dataset in order to apply a linear regression model to a reduced number of variables while retaining the bulk of the original predictor's variability.

G. Artificial Neural Network

Classification, prediction, and association problems have all been solved using artificial neural networks (ANNs). Nonlinear mathematical functions may be approximated by ANNs, which is especially useful when the connection between the variables is unknown or complicated. Among the feed-forward networks investigated in this study was the Multilayer Perceptron (MLP), which can hire backpropagation, conjugate gradient, and other methods. The diagram above depicts the MLP ANN architecture. To begin with, in an ANN's input layer [8], there is an array of input units (x_i, i =1, 2, n) and random weights (w_i), which are typically in the range of [-1,1]. [9] Every hidden (middle) layer unit receives this information while calculating the weighted average of all x_i values. The output of the hidden layer, denoted as y_c, is computed by summing the inputs multiplied by their weights, as illustrated in the equation:

$$y_c = f \sum_{i=1}^{n} w_i x_i \qquad (3)$$

where f is the user-selected activation function (sigmoid, tangent hyperbolic, exponential, linear, step, or other).

In our case, we have used a 10-layered Deep [10] feed-forward neural network with the rectified linear unit as the activation function. The challenge at hand determines the number of input and output neurons. There are half as many

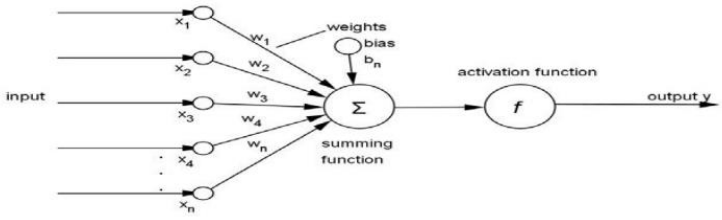

Fig. 3. ANN Architecture

neurons in each succeeding layer as there were in the input layer (1024 neurons), which is how these networks were trained, tested, and verified. The AQI value predicted by the output neuron (layer) is the most accurate.

H. Applying PCA in conjunction with Artificial Neural Networks (PC-ANN)

Using PCA and a neural network to improve the model's performance is the ultimate goal of this strategy. [11]. PCA is a multivariate statistical method that is frequently used to analyze air pollution data. The goal of PCA is to minimize the number of predictive elements and transform them into new variables known as principal components. These additional variables are formed by merging original data into different linear combinations with the greatest possible variety. PCA also reduces data set collinearity, which leads to the worst air pollution concentration forecasts (Sousa et al., 2007) [12]. Since a PCA neural network requires fewer input data and input variables than a neural network, it is easier to create and run than a neural network. The correlation matrix of the normalized input data may be used to calculate the PCs. Using the characteristic equation below, we can get the eigenvalues of the correlation matrix "C".

$$|C - \lambda I| = 0 \tag{4}$$

The eigenvalue is λ, and the identity matrix is I. A non-zero eigenvector e exists for each eigenvalue λ, which is defined as:

$$Ce = \lambda e \tag{5}$$

The correlation matrix yields the eigenvectors, which have mutually orthogonal linear combinations. The entire number of variances explained by each of the eigenvectors is represented by their associated eigenvalues. A significant portion of the overall variance can be explained by keeping the top few pairs of eigenvalues–eigenvector or principal components. It's able to categorize noise as higher-order primary components that only make up a small percentage of the total variance. The ith PC's variance is stated as:

$$Variance_i = \frac{\hat{\lambda}_1}{\sum_n \lambda_n} \tag{6}$$

The linear combination of factors (the eigenvalue of PC1) accounts for a significant portion of the total variability in the data. The PC1 does not take into consideration the greatest amount of unpredictability, which is accounted for by PC2. A similar trend is seen in the PCs that preceded this one. Multiplying eigenvectors [13] with the original data set yields the orthogonal set, which is then used to locate the PCs.

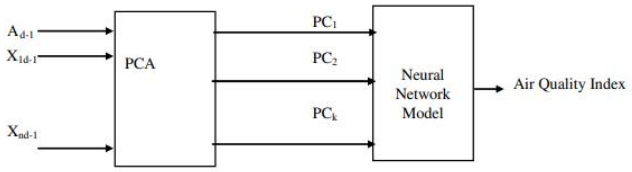

Fig. 4. PC-ANN Architecture

I. Model Performance Determination

The MLR, PC-ANN, and PCR models' overall utility and performance were scored according to one of four distinct criteria in order to provide an accurate prediction. The following are the criteria that were used in the selection process: Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). In this case, a lower RMSE, MAPE, MAE, and a higher R2 score result in a more accurate prediction.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i)^2}{\sum_{i=1}^{n} (y_i - y_i)^2} \tag{7}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{8}$$

$$\text{MAPE} = \left(\frac{\sum_{i=1}^{n} \left| y_i - \hat{y}^i \right| / y_i}{n} \right) * 100 \tag{9}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_1|}{n} \tag{10}$$

IV. Results And discussions

A. Principal Component Analysis

To identify the number of PCs to be used to get accurate predictions and results usually the variance explained by each of the PCs is analyzed, the higher the cumulative variance explained by the number of PCs selected better will be the results, in our case, the first 8 PCs, which accounted for nearly 99% of the total variation in the data, had to be chosen. The remaining higher-order PCs are considered noise and are discarded. A Plot explaining the cumulative variance by all the PCs has been presented in the figure below.

B. Using MLR and PCR Models to Predict the AQI

A model based on various linear regression (MLR) is used to estimate the AQI for a particular location in this study, with the quality of the air and a variety of environmental characteristics (11 parameters) serving both as independent and dependent variables. A two-step approach is used to calculate PCR. The data is first analyzed using a principal

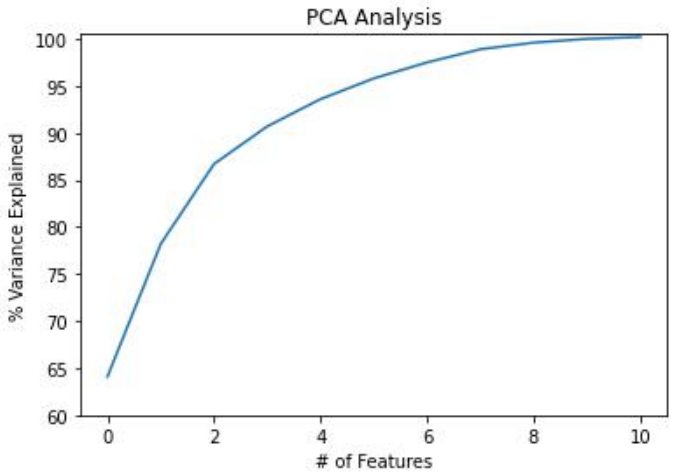

Fig. 5. Percentage Variance vs Number of Features

component analysis (PCA). After that, the eight PCs that produced the data are used as MLR predictors. To reduce the number of variables to predict, PCR may highlight the data set's linear structure.

C. ANN and PC-ANN Models for AQI Prediction

For the ANN model, 11 original meteorological and pollution factors are used as input, and the designed ANN has a total of 8 hidden layers, For the PC-ANN model the original parameters are reduced to 8 principal components using principal component analysis, An ANN model with 8 hidden layers and a single neuron in the output layer is then used to predict AQI using these eight PCs as input [14].

D. Comparative performance of ANN, MLR, PCR and PC-ANN models

Analytical methods use data from eleven air pollutant and meteorological parameters total (the original data). Principal Component Analysis (PCA) is applied to condense the initial eleven parameters into eight PCs for the remaining procedures (PCA). ANN and MLR models both employ the eight produced PCs as input variables[15]. The model's quality was evaluated using RMSE, MAPE, MAE, and R2. The table below shows the comparison of the ANN, MLR, PCR, and PC-ANN models.

TABLE II
MODEL PERFORMANCE TABLE

Model	R2	RMSE	MAE	MAPE
MLR	0.88641	42.8000	30.5265	0.20237
ANN	0.99444	8.9479	4.9945	0.02320
PCR	0.88644	41.5212	30.6909	0.21504
PC-ANN	0.99691	6.9489	4.0988	0.02531

The PC-ANN model outperforms all other models in terms of AQI estimate, as shown in the data above. It can also be seen that in general, models which have been incorporated along principal component analysis are performing better than their base models (PCR outperforms MLR also PC-ANN outperforms ANN).

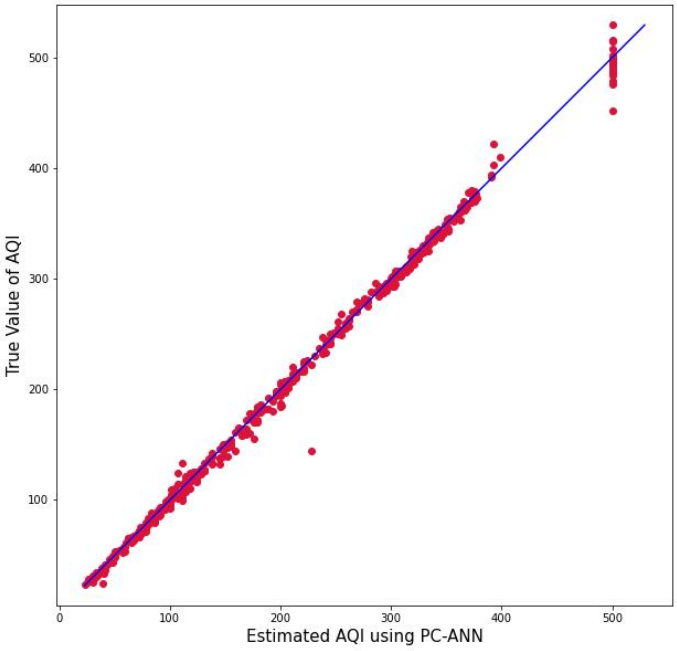

Fig. 6. Estimated AQI using PC-ANN method Vs Actual value of AQI

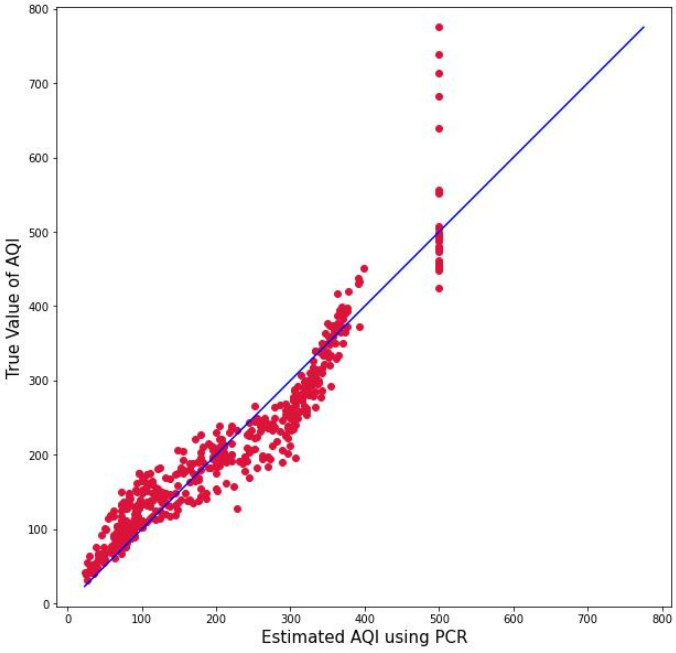

Fig. 7. Estimated AQI using PCR method Vs Actual value of AQI

V. CONCLUSIONS

The fundamental goal of this research is to predict the AQI using a mix of PCA and artificial neural networks (PC-ANN). Some of the models that are used include MLR, ANN, PC-ANN, and PCR. Models that use principal components

analysis to lower the original parameters are called PCR, while models that use ANN and MLR use real raw data.

The results show that using Eleven original parameters as inputs, the ANN model produces a high R2 and low error rates, in contrast to the MLR model's outputs. However, the PC-ANN model surpasses the ANN, MLR, and PCR models in terms of R2 value and error rate values.

In predicting the results of a study, an ANN or PC-ANN model performs best than MLR or PCR. Even more importantly, this study demonstrates how valuable neural networks and PC-ANN models can be for solving local pollution control issues.

REFERENCES

[1] I. Ungvári et al., "Relationship between air pollution, NFE2L2 gene polymorphisms and childhood asthma in a Hungarian population," Journal of community genetics, vol. 3, no. 1, pp. 25–33, 2012.

[2] J. Kotcher, E. Maibach, and W.T. Choi, "Fossil fuels are harming our brains: identifying key messages about the health effects of air pollution from fossil fuels," BMC public health, vol. 19, no. 1, p. 1079, 2019.

[3] Kumar A., Goyal P. (2011), Forecasting of air quality index in Delhi using principal component regression technique, atmospheric pollution research 2, 436–444.

[4] Khan R. A., Zain Sh. M., Juahir H., Yusoff M. K. and Tg Hanidza T. I.; using principal component scores and artificial neural networks in predicting water quality index; chemometrics in practical applications edited by Dr. Kurt Varmuza, Chapter 12, 2012.

[5] Sarwat, Eman & El-shanshoury, Ghada. (2018). Estimation of air quality index by merging neural network with principal component analysis. International journal of computer application, 8.10.26808/rs.ca.v8n1.01.

[6] Van den elshout S., Leger K., Fabio N. (2008), Comparing urban air quality in Europe in real time: A review of existing air quality indices and the proposal of a common alternative, Environment international 34, 720–726.

[7] "Emerging research in data engineering Systems and computer communications" ,springer science and business media LLC, 2020.

[8] Marijana Zekić-Sušac, Sanja Pfeifer, Nataša Šarlija. "A comparison of machine learning methods in a high-dimensional classification problem", Business systems research journal, 2014.

[9] Ljiljana Majnaric, Marijana Zekic-Susac. "Elucidating clinical context of lymphopenia by nonlinear modelling", Expert Systems with Applications, 2012.

[10] Azid A., Juahir H., Latif M. T., Zain Sh. M. and Osman M. R.; Feed-forward artificial neural network model for air pollutant index prediction in the southern region of peninsular Malaysia; journal of environmental protection, No. 4, pp.1-10 , 2013.

[11] Kumar A., Goyal P. (2012), Forecasting of air quality index in Delhi using neural network based on principal component analysis Pure Appl. Geophys. 170 (2013), 711–722.

[12] Sousa S. I. V., Martins, F. G. Alvim-Ferraz M. C. M., Pereira M.C. (2007), Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, Environmental modelling software 22,97–103.

[13] Dhirendra Mishra, Pramila Goyal. "Development of artificial intelligence based NO 2 forecasting models at Taj Mahal, Agra" ,Atmospheric pollution research, 2015.

[14] Nagendra S. M. S., Venugopal K., Jones S. L. (2007), Assessment of air quality near traffic intersections in Bangalore city using air quality index, Transportation research part D 12, 167–176.

[15] Cho, K.H.. "Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network",Water Research, 20111101.